2-2005

Protein sequence entropy is closely related to packing density and hydrophobicity

H. Liao
San Jose State University

W. Yeh
San Jose State University

D. Chiang
Sage-N Research

Robert L. Jernigan
Iowa State University, jernigan@iastate.edu

B. Lustig
San Jose State University

Protein sequence entropy is closely related to packing density and hydrophobicity

Abstract

We investigated the correlation between the Shannon information entropy, 'sequence entropy', with respect to the local flexibility of native globular proteins as described by inverse packing density. These are determined at each residue position for a total set of 130 query proteins, where sequence entropies are calculated from each set of aligned residues. For the accompanying aggregate set of 130 alignments, a strong linear correlation is observed between the calculated sequence entropy and the corresponding inverse packing density determined at an associated residue position. This region of linearity spans the range of Cα packing densities from 12 to 25 amino acids within a sphere of 9 Å radius. Three different hydrophobicity scales all mimic the behavior of the sequence entropies. This confirms the idea that the ability to accommodate mutations is strongly dependent on the available space and on the propensity for each amino acid type to be buried. Future applications of these types of methods may prove useful in identifying both core and flexible residues within a protein.

Keywords

hydrophobicity, sequence entropy, sequence-structure relationship, sequence variability

Disciplines

Biochemistry | Biophysics | Computational Biology | Structural Biology

Comments

This is a manuscript of an article published as Liao, H., W. Yeh, D. Chiang, R. L. Jernigan, and Brooke Lustig. "Protein sequence entropy is closely related to packing density and hydrophobicity." *Protein Engineering Design and Selection* 18, no. 2 (2005): 59-64. doi:10.1093/protein/gzi009. Posted with permission.

Published in final edited form as:
Protein Eng Des Sel. 2005 February ; 18(2): 59–64. doi:10.1093/protein/gzi009.

Protein sequence entropy is closely related to packing density and hydrophobicity

H. Liao[1], **W. Yeh**[2], **D. Chiang**[3], **R.L. Jernigan**[4], and **B. Lustig**[1,5]

1 *Department of Chemistry, San Jose State University, San Jose, CA 95192-0101*

2 *Department of General Engineering, San Jose State University, San Jose, CA 95192-0101*

3 *Sage-N Research, Saratoga, CA 95070-6082*

4 *L.H. Baker Center for Bioinformatics and Biological Statistics, Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA 50014, USA*

Abstract

We investigated the correlation between the Shannon information entropy, 'sequence entropy', with respect to the local flexibility of native globular proteins as described by inverse packing density. These are determined at each residue position for a total set of 130 query proteins, where sequence entropies are calculated from each set of aligned residues. For the accompanying aggregate set of 130 alignments, a strong linear correlation is observed between the calculated sequence entropy and the corresponding inverse packing density determined at an associated residue position. This region of linearity spans the range of C^{α} packing densities from 12 to 25 amino acids within a sphere of 9 Å radius. Three different hydrophobicity scales all mimic the behavior of the sequence entropies. This confirms the idea that the ability to accommodate mutations is strongly dependent on the available space and on the propensity for each amino acid type to be buried. Future applications of these types of methods may prove useful in identifying both core and flexible residues within a protein.

Keywords

hydrophobicity; sequence entropy; sequence-structure relationship; sequence variability

Introduction

General studies of the geometries within proteins have a long history and have lead to important insights into protein structure (Chothia *et al*., 1981; Chothia and Finkelstein, 1990; Maritan *et al*., 2000; Banavar *et al*., 2002). Specific studies of the packing geometries have indicated, for coarse-grained structures with one point per residue, that amino acids pack in local clusters with the same orientations as close-packed spheres (Bagci *et al*., 2002, 2003). At the same time, cavities within protein structures are known to be important for function (Doyle *et al*., 1998; Sigler *et al*., 1998; Zhang *et al*., 2003).

Globular proteins are compact and hence densely packed (Richards, 1974), even to the extent that their interior is frequently viewed as being solid-like (Hermans and Scheraga, 1961; Richards, 1997); however, there are still numerous voids and cavities in protein interiors (Liang

5To whom correspondence should be addressed. E-mail: blustig@science.sjsu.edu.

Edited by Harold Scheraga

and Dill, 2001). The importance of tight packing is widely acknowledged and is thought to be important for protein stability (Ericksson *et al*., 1992; Privalov, 1996), for nucleation of protein folding (Ptitsyn, 1998; Ptitsyn and Ting, 1999; Ting and Jernigan, 2002) and for the design of novel proteins (Dahiyat and Mayo, 1997). In conjunction with nucleation, it has previously been posited that the conservation of amino acid residues through evolution may include essential tightly packed sites (Mirny *et al*., 1998; Ptitsyn, 1998; Ptitsyn and Ting, 1999; Ting and Jernigan, 2002).

However, the exact relationship between sequence and structure is only partially understood (Jones, 2000; Baker and Sali, 2001), which is the subject of this paper. Whereas protein sequence is easily determined, 3-D structure is significantly more difficult. Employing sequence alignments in conjunction with molecular modeling has proven to be among the most successful computational methodologies for protein structure prediction (Bryant and Lawrence, 1993; Marti-Renom *et al*., 2000). One key assumption in homology-based modeling is that conserved regions share structural similarities, but the structural basis of this connection has not been clearly determined.

Multiple alignments of regions of secondary structure may be useful in the identification of key hydrophobic residues when utilizing hydrophobic cluster analysis (Poupon and Mornon, 1999; Gross *et al*., 2000). Determining patterns of variability within amino acid sequence by using information theory has also proven useful in identifying unique protein secondary structures (Pilpel and Lancet, 1999). Large-scale exploration of sequence space has shown clustering of sequence entropy values corresponding to a particular fold (Larson *et al*., 2002). The application of Shannon entropy to nucleic acid sequence variability has proven to be a useful tool in identifying control regions in DNA (Schneider *et al*., 1986) and has been extended as one of several methods of scoring amino acid conservation in proteins (Zou and Saven, 2000; Valdar, 2002).

Shannon entropies for protein sequence have been shown to correlate with entropies calculated from local physical parameters, including backbone geometry (Koehl and Levitt, 2002). Interestingly, conventional generalized chain statistics appear to overweigh significantly the magnitude of the entropic penalty associated with loop closure in proteins and RNA (Lustig *et al*., 1998; Scalley-Kim *et al*., 2003). It is clear that continued exploration of the connections between entropy, structure and sequence is critical to a better understanding of protein stability and function.

Although there have been some demonstrations of connections between sequence conservation and structural properties (Demirel *et al*., 1998), there are no definitive studies on this subject. Establishing direct connections between sequences and structural features has proven difficult, hence the limited number of successes at protein design and the limited understanding of mutagenesis. Recent applications of sequence variability to structure predictions have enhanced results, so empirical measures of sequence variability are useful by themselves, even if their full implications are not well understood in terms of structural features.

While investigations of packing of protein atoms would likely be informative, we chose here to investigate coarse-grained packing among points each representing a neighboring amino acid. The results we will see are then more general, even if not so directly useful in predictions related to protein design.

Here we generate a large set of aligned protein sequences generated from a diverse sample of 130 protein sequences. Sequence entropies for individual residues are calculated. They are then compared with the corresponding local flexibility as measured by the extent of C^{α} packing calculated from the corresponding structures. Similar comparisons are also made between the residue hydrophobicity and the corresponding packing.

Methods

A diverse, well-characterized set of 130 protein sequences (Table I) was compiled from the Protein Data Bank (2002). Redundant proteins were removed. Sequences are utilized from a wide variety of proteins including multi-chain proteins, where 18% involve multi-chain proteins and the remaining 107 sequences are single-chain proteins. Aligned sequences are generated against each of these protein sequences, with BLASTP (Altschul *et al*., 1997) searching GenBank as available from the National Center of Biotechnology Information (2002). Alignments are not included if bit scores fall below 100 and they must be at a level ⩾40% of the best score. Calculations with a representative set of proteins showed 40% of the BLASTP bit score as a reasonable threshold with respect to calculations of sequence entropy and their dependence on density.

Also at least 10 sequences are required. A maximum number of 100 alignments is typically allowed. The result generates a representative distribution of 7143 aligned protein sequences. The average and median number of alignments per query and the overall range of numbers of alignments are 55, 55 and 10–100, respectively. The frequency distribution of the BLASTP bit scores for all 130 sets of alignments is consistent with the right-skewed (i.e. positive skew) distribution for a randomized set of BLAST scores (Altschul *et al*., 1994). Here the mean, median and the overall range of BLASTP bit scores for all 7143 alignments are 408, 354 and 100–1793, respectively.

For protein sequences an expression for sequence entropy S_k at amino acid position k is expressed as

$$S_k = -\sum_{j=1,20} P_{jk} \ln P_{jk} \tag{1}$$

where the probability P_{jk} at some amino acid sequence position k is derived from the frequency f_{jk} for an amino acid type j at sequence position k for all of the aligned residues. Although gaps could have been assigned as an additional amino acid type, we chose to ignore them here. In order to compare against the random case, we subtract the following term (Gerstein and Altman, 1995) from Equation 1:

$$S_k^R = -\sum_{j=1,20} P_j \ln P_j \tag{2}$$

where P_j is the probability of amino acid type j over all alignments.

For each residue from the 130 sample protein sequences, C^α packing densities are calculated using their associated atomic coordinates. An optimal radius of C^α packing was determined for 9 Å around a given C^α residue position. In limited preliminary investigations this value was found to be best; greater scatter is observed for example in the single average entropies for radii of 10 and 11 Å. Smaller values omit some important cases in the distribution. Here we investigate the extent to which the inverse of the local packing density, as a measure of local flexibility (Bahar *et al*., 1997), is correlated with sequence variability.

Results

Calculated sequence entropy (Equation 1) for each protein is compared against the inverse C^α packing density (see Table I for summary). Typically, the probability P that the observed data could come from a randomized population (Bevington, 1969) for individual proteins falls below 0.001. A selection of correlation plots are shown in Figure 1A, B and C for pepsinogen (3psg, 365 aligned residues), dihydrofolate reductase (4dfr, 158 aligned residues) and oncomodulin (1omd, 107 aligned residues), respectively. The respective slopes are 13.020,

6.064 and 4.328, with respective correlation coefficients 0.447, 0.274 and 0.141. Data were collected in bins for each integral number of residues falling within a sphere of per 9 Å radius. For most single protein correlation plots the slopes remain effectively unchanged upon averaging.

In total, there are 41 543 query residues following the removal of the 89 extreme outlying values indicated outside the two arrows shown in Figure 2. The mean and median frequency values per density interval of one C^{α} per 9 Å radius are unchanged at 14.6 and 15. The overall (i.e. for all 130 alignment sets) sequence entropy versus inverse C^{α} packing density correlation plots are shown in Figure 3A. Here, a single average is performed by summing individual residue entropies for a particular C^{α} packing density interval from all 130 sets of protein alignments. 'Double' averaging entails first averaging the entropy per density interval for individual proteins, before averaging over the full set of proteins. Except for a significant reduction in standard deviations with the 'double' averaging procedure, the two types of averaged sequence entropy are essentially identical.

There are two major regions corresponding to high and low densities observable in the correlation plots of sequence entropy versus inverse packing density in Figure 3A. Note that a similar overall pattern of single averaged sequence entropy was observed when the effects of randomness were accounted for by subtracting the term shown in Equation 2. Region I, with a steep slope, corresponds to the higher packing densities of 25 to 12 C^{α} atoms (inverse density from 0.040 to 0.083), where an increase in sequence entropy is clearly proportional to the inverse density. Region II to the right still includes a significant number of residues (10 173) and is found to be nearly constant in calculated sequence entropy, involving packing densities ranging from 11 to 6 (representing an upper bound inverse density of 0.17). It is logical that beyond a certain packing density, changes in sequence entropy remain uncorrelated.

Region I, in the overall correlation plots (Figure 3B), involves 74.9% of all the sample protein residues. Here the single averaged and 'double' averaged sequence entropies are shown to be strongly linearly correlated with the inverse packing density. The straight-line fit for the single averaged sequence entropy versus inverse packing density is $y = 12.350x - 0.20$; the correlation coefficient is 0.997; $P < 0.001$. The straight-line fit involving the 'double' averaged entropy is effectively identical. Region II, accounting for an additional 24.4% of the sample protein residues, indicates for strongly hydrophobic residue types (Poupon and Mornon, 1999) an apparent limiting fraction (Figure 3A) of about 10%. This suggests a threshold for the number of hydrophobic residues embedded in regions that are probably accessible to water.

Shown in Figure 4 A is a superposition of normalized averaged sample protein hydrophobicities and single averaged sequence entropy, as a function of inverse packing density. Using three different scales (Hopp and Woods, 1981;Engelman *et al*., 1986;Sharp *et al*., 1991), hydrophobicity is calculated for every query protein residue that is part of an alignment. For Hopp and Woods (1981) calculations by Levitt (1976) were also included. With each scale, a normalized hydrophobicity is calculated for the set of all residues within a density interval. Then those three normalized hydrophobicity plots (see Figure 4B) are averaged and renormalized again. Superimposed is the smooth curve normalized representation (determined from original values in Figure 3A) of values for sequence entropy. Clearly, all three sets of hydrophobicity values, calculated for each scale (Figure 4B), resemble the corresponding sequence entropy values.

Discussion

Flexibility and sequence entropy

Previously a strong correlation has been reported between computed displacements based on elastic networks reflecting residue packing (Bahar *et al*., 1998) and measured hydrogen exchange (HX). The freedom to move a residue is entropic in character. Regions of high packing density resist hydrogen exchange, because of both stability and inaccessibility. Here, we have gone further to relate our calculated inverse C^{α} packing density from X-ray structures to the sequence variabilities. Strong linear correlations are observed between sequence entropy and the inverse packing density, except at the highest and low ranges of densities. This provides a quantitative relationship between these two quantities and an important structural measure for determining likely sites for mutagenesis.

The selection of sequences to be included in sequence analysis is a difficult problem and results can depend strongly on the selection procedure. Ptitsyn (1998) advocated selection of conserved clusters of sequence sets determined by including only distantly related species. However, here we simply used the sequence matches from GenBank without any filtering. Despite this, the overall trends are extremely clear, although to a limited extent within individual proteins.

In addition, the correlation between sequence variability and motility is consistent with a similar pattern that we noted with respect to peptide binding to RNA (Hsieh *et al*., 2002). Enhanced motility at a particular residue position is associated with the ability of local structure to accommodate mutation. Such behavior can more broadly be related to sequence variability in a folded protein. The ability to accommodate mutations corresponds to allowing a range of positions, including possible contacts.

Hydrophobicity and sequence variability

The strong correlation between calculated sequence entropy and the hydrophobicity shown in Figure 4 is remarkable. For each protein, its sequence entropy is calculated at each sequence position. This simply reflects the sequence variability at that position. The hydrophobicity for each residue position of each original single sequence is averaged for each bin over just the 130 sample proteins. It is important to remember that here the sequence entropy and the hydrophobicity calculations are both averaged over all residues within each density bin. In addition, the three sets of hydrophobicity scales (Hopp and Woods, 1981;Engelman *et al*., 1986;Sharp *et al*., 1991) are diverse in their origins and include experimental optimization and/or validation based on a variety of systems. Calculations by Levitt (1976) were also included for use by Hopp and Woods (1981). The lack of any significant differences among the three sets of normalized hydrophobicity values (Figure 4B) as a function of inverse density suggests that the relative differences among individual amino acids within a hydrophobicity scale are largely compensated among other values within that set. Clearly, correlations between the sequence variabilities reflected in the sequence entropies and the corresponding hydrophobicities are consistent with the average behavior for residues with a given packing density. Still, this observed correlation between average sequence entropy and hydrophobicity is remarkable, but both are reflecting fundamental properties relating to the extent of burial. The critical importance of hydrophobicity for folding of model protein chains (Hinds and Levitt, 1994;Dill *et al*., 1995) is well known. This is consistent with the fact that key hydrophobic residues can be described as buried or tightly packed (Ptitsyn, 1998;Ting and Jernigan, 2002).

Packing and the resulting interactions associated with hydrophobicity are not a simple matter of just accounting for pairs of contacts (Dima and Thirumalai, 2004). In packing multiple

contacts are usual. Our calculation of C^{α} packing density represents a coarse-grained counting of such contacts, but is a less detailed consideration. We show that the local flexibility is closely related to the inverse of the coarse-grained packing density.

Here, sequence variability as measured by sequence entropy is correlated with the inverse of the residue packing. The propensity for packing of a particular amino acid type reflects its hydrophobicity and side chain entropy (Pickett and Sternberg, 1993). Notably, average contact energies for the various amino acid pairs also correlate well with existing hydrophobicity scales (Young *et al*., 1994). This suggests that in principle these are strongly entropic in nature. It might be possible to calculate more directly configurational entropies in lieu of the comparable inverse density measure of relative flexibility, by using full atomic representation. Such calculations would depend upon a residue's environment in more realistic ways than given by simple residue density. T his might also reduce the range for individual residue entropies calculated from sequence variability within a density bin.

Progress in this direction would assist with protein design, a closely related problem (Dahiyat and Mayo, 1997; Li *et al*., 1998; Buchler and Goldstein, 1999; Shih *et al*., 2000; Tiana *et al*., 2001; Koehl and Levitt, 2002; Larson *et al*., 2002; England *et al*., 2003). Further studies in the direction of the present work could lead to better predictions of sustainable sequence substitutions. However, from the present results it appears that every measure of packing density for single residues of a single protein does not necessarily correlate well with the sequence conservation at that site. Further efforts are clearly required to achieve this goal; however, the present results begin to point out a way for achieving such a goal.

Conclusion

Here packing at the residue level for coarse-grained structures has been shown to exhibit a strong connection to sequence conservation, by the practice of averaging over large numbers of residues. Why is this averaging necessary? One possible explanation is that the large number of combinations of ways in which a residue's atoms can be packed together requires averaging over large numbers of occurrences, in order to obtain a meaningful single representation of all these combinations. It is also possible that residue size may affect the results, so that averaging over many occurrences will fully account for all of the various types of neighboring residues including individual side chain conformations.

Two distinct behaviors are identified for different inverse packing density regions (Figures 3 and 4). In the first region, 74.9% of sequence positions exhibit a linear dependence of sequence entropy over the inverse C^{α} packing density range 0.040–0.083, whereas in the second region, having inverse packing density >0.083, another 24.4% of query positions typically indicate a nearly constant sequence entropy. This saturation suggests that up to a certain minimum number of residues are allowed in low-density regions. Moreover, a certain fraction of those residues are hydrophobic and would appear to be accessible to water, consistent with a considerable lack of restrictions on the types of residues that can be accommodated. All of this suggests that for most residue positions the ability to accommodate sequence substitutions as measured by sequence entropy is inversely correlated with the extent of their packing. Also, on average for a particular amino acid type, hydrophobicity is correlated with the degree of residue packing. Deeper understanding of the connections between structural properties and sequence entropy awaits further study. However, the future development of such sequence entropy methods for the identification of core as well as flexible residues appears promising.

Acknowledgements

This work included resources from BERI (Biotechnology Education and Research Institute) at San Jose State University and CSUPERB (California State University Program for Education and Research in Biotechnology).

References

Altschul SF, Boguski MS, Gish W, Wooten JC. Nat Genet 1994;6:119–129. [PubMed: 8162065]

Altschul SF, Madden TL, Scaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Nucleic Acids Res 1997;25:3389–3402. [PubMed: 9254694]

Bagci Z, Jernigan RL, Bahar I. J Chem Phys 2002;116:2269–2276.

Bagci Z, Kloczkowski A, Jernigan RL, Bahar I. Proteins 2003;53:56–67. [PubMed: 12945049]

Bahar I, Atilgan AR, Erman B. Fold Des 1997;2:173–181. [PubMed: 9218955]

Bahar I, Wallqvist A, Covell DG, Jernigan RL. Biochemistry 1998;37:1067–1075. [PubMed: 9454598]

Baker D, Sali A. Science 2001;294:93–96. [PubMed: 11588250]

Banavar JR, Maritan A, Micheletti C, Trovato A. Proteins 2002;47:315–322. [PubMed: 11948785]

Bevington, PR. Data Reduction and Error Analysis for the Physical Sciences. McGraw-Hill; New York: 1969. p. Appendix C

Bryant SH, Lawrence CE. Proteins 1993;16:92–112. [PubMed: 8497488]

Buchler NE, Goldstein RA. Proteins 1999;34:113–124. [PubMed: 10336377]

Chothia C, Finkelstein AV. Annu Rev Biochem 1990;59:1007–1039. [PubMed: 2197975]

Chothia C, Levitt M, Richardson D. J Mol Biol 1981;145:215–250. [PubMed: 7265198]

Dahiyat BI, Mayo SL. Science 1997;278:82–87. [PubMed: 9311930]

Demirel MC, Atilgan AR, Jernigan RL, Erman B, Bahar I. Protein Sci 1998;7:2522–2532. [PubMed: 9865946]

Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, Thomas PD, Chan HS. Protein Sci 1995;4:561–602. [PubMed: 7613459]

Dima RI, Thirmalai D. J Phys Chem B 2004;108:6564–6570.

Doyle DA, Cabral JM, Pfuetzner RM, Kuo A, Gulbis JM, Cohen SL, Chait BT, MacKinnon R. Science 1998;280:69–77. [PubMed: 9525859]

Engelman DM, Steitz TA, Goldman A. Annu Rev Biophys Biophys Chem 1986;15:321–353. [PubMed: 3521657]

England JL, Shakhnovich BE, Shakhnovich EI. Proc Natl Acad Sci USA 2003;100:8727–8731. [PubMed: 12843403]

Eriksson AE, Baase WA, Zhang XJ, Heinz DW, Blaber M, Baldwin EP, Matthews BW. Science 1992;255:178–183. [PubMed: 1553543]

Gerstein M, Altman RB. J Mol Biol 1995;251:161–175. [PubMed: 7643385]

Gross EA, Li GR, Lin ZY, Ruuska SE, Boatright JH, Mian IS, Nickerson JM. Mol Vis 2000;6:30–39. [PubMed: 10756179]

Hermans J, Scheraga HA. J Am Chem Soc 1961;83:3293–3330.

Hinds DA, Levitt M. J Mol Biol 1994;243:668–682. [PubMed: 7966290]

Hopp TP, Woods KR. Proc Natl Acad Sci USA 1981;78:3824–3828. [PubMed: 6167991]

Hsieh M, Collins ED, Blomquist T, Lustig B. J Biomol Struct Dyn 2002;20:243–251. [PubMed: 12354076]

Jones DT. Curr Opin Struct Biol 2000;10:371–379. [PubMed: 10851185]

Koehl P, Levitt M. Proc Natl Acad Sci USA 2002;99:1280–1285. [PubMed: 11805293]

Larson SM, England JL, Desjarlais JR, Pande VS. Protein Sci 2002;11:2804–2813. [PubMed: 12441379]

Levitt M. J Mol Biol 1976;104:59–107. [PubMed: 957439]

Li H, Tang C, Wingreen NS. Proc Natl Acad Sci USA 1998;95:4987–4990. [PubMed: 9560215]

Liang J, Dill KA. Biophys J 2001;81:751–766. [PubMed: 11463623]

Lustig B, Bahar I, Jernigan RL. Nucleic Acids Res 1998;26:5212–5217. [PubMed: 9801321]

Maritan A, Micheletti C, Trovato A, Banavar JR. Nature 2000;406:287–290. [PubMed: 10917526]

Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Annu Rev Biophys Biomol Struct 2000;29:291–325. [PubMed: 10940251]

Mirny L, Abkevich VL, Shakhnovich EI. Proc Natl Acad Sci USA 1998;95:4976–4981. [PubMed: 9560213]

National Center for Biotechnology Information. 2002. http://www.ncbi.nlm.nih.gov/

Pickett SD, Sternberg MJE. J Mol Biol 1993;231:825–839. [PubMed: 8515453]

Pilpel Y, Lancet D. Protein Sci 1999;8:969–977. [PubMed: 10338007]

Poupon A, Mornon JP. Theor Chem Acc 1999;101:2–8.

Privalov PL. J Mol Biol 1996;258:707–725. [PubMed: 8637003]

Protein Data Bank. 2002. http//www.rcsb.org.pdb/

Ptitsyn OB. J Mol Biol 1998;278:655–666. [PubMed: 9600846]

Ptitsyn OB, Ting KL. J Mol Biol 1999;291:671–682. [PubMed: 10448045]

Richards FM. J Mol Biol 1974;82:1–14. [PubMed: 4818482]

Richards FM. Cell Mol Life Sci 1997;53:790–802. [PubMed: 9413550]

Scalley-Kim M, Minard P, Baker D. Protein Sci 2003;12:197–206. [PubMed: 12538883]

Schneider TD, Stormo GD, Gold L. J Mol Biol 1986;188:415–431. [PubMed: 3525846]

Sharp KA, Nicholls A, Friedman R, Honig B. Biochemistry 1991;30:9686–9697. [PubMed: 1911756]

Shih CT, Su ZY, Gwan JF, Hao BL, Hsieh CH, Lee HC. Phys Rev Lett 2000;84:386–389. [PubMed: 11015917]

Sigler PB, Xu Z, Rye HS, Burston SG, Fenton WA, Horwich AL. Annu Rev Biochem 1998;67:581–608. [PubMed: 9759498]

Tiana G, Broglia RA, Provasi D. Phys Rev E 2001;64:011904_1–6.

Ting KL, Jernigan RL. J Mol Evol 2002;54:425–436. [PubMed: 11956682]

Valdar WSJ. Proteins 2002;48:227–241. [PubMed: 12112692]

Young L, Jernigan RL, Covell DG. Protein Sci 1994;3:717–729. [PubMed: 8061602]

Zhang J, Chen R, Tang C, Liang J. J Chem Phys 2003;118:6102–6109.

Zou J, Saven JG. J Mol Biol 2000;296:281–294. [PubMed: 10656832]

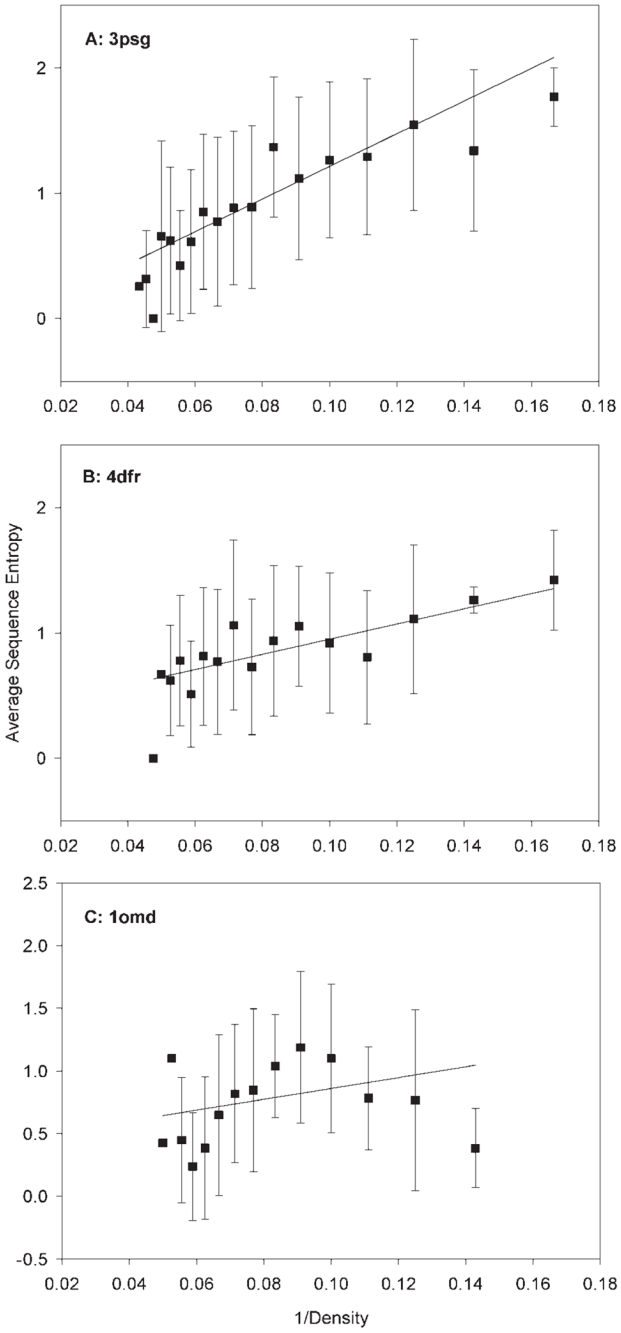

Fig. 1.

Correlation between sequence entropy and inverse of packing density for a range of proteins. The inverse of packing density (abscissa) is calculated from the sample protein's atomic coordinates, determining the number of residue's C^{α} atoms within a 9 Å radius. Sequence entropy is calculated from a sequence alignment set generated by BLASTP from the query sequence. Average entropy (closed squares) is also determined by averaging the sequence entropy for all sequence positions falling within an interval of packing density. Error bars corresponding to standard deviation calculated from the data and the linear fit for all points (line) are shown. (**A**) For pepsinogen (3psg: 365 aligned residues), the straight-line fit for all data is $y = 13.020x - 0.09$ with correlation coefficient 0.447 and $P < 0.001$. For averaged data

$y = 12.070x - 0.09$ with correlation coefficient 0.898 and $P < 0.001$. (**B**) For dihydrofolate reductase (4dfr: 158 aligned residues), the straight-line fit for all data is $y = 6.064x + 0.34$ with correlation coefficient 0.274 and $P < 0.001$. For averaged data $y = 7.350x + 0.22$ with correlation coefficient 0.796 and $P < 0.001$. (**C**) For oncomodulin (1omd: 107 aligned residues), the straight-line fit for all data is $y = 4.328x + 0.43$ with correlation coefficient 0.141 and $P < 0.15$. For averaged data $y = 1.624x + 0.59$ with correlation coefficient 0.149 and $P < 0.15$.

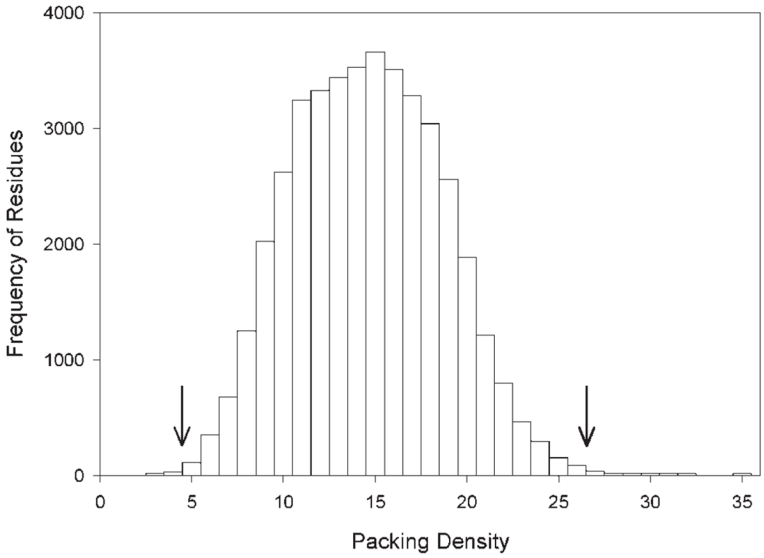

Fig. 2.

Frequency distribution of the number of aligned residues as a function of C^{α} density within a radius of 9 Å. The total original 41 632 query protein residues for the set of 130 proteins have a mean packing density of 14.6, a median of 15 and SD 4.056. These values remain effectively unchanged for the 41 543 residues remaining following the removal of outlying values to the left of the first arrow and to the right of the second.

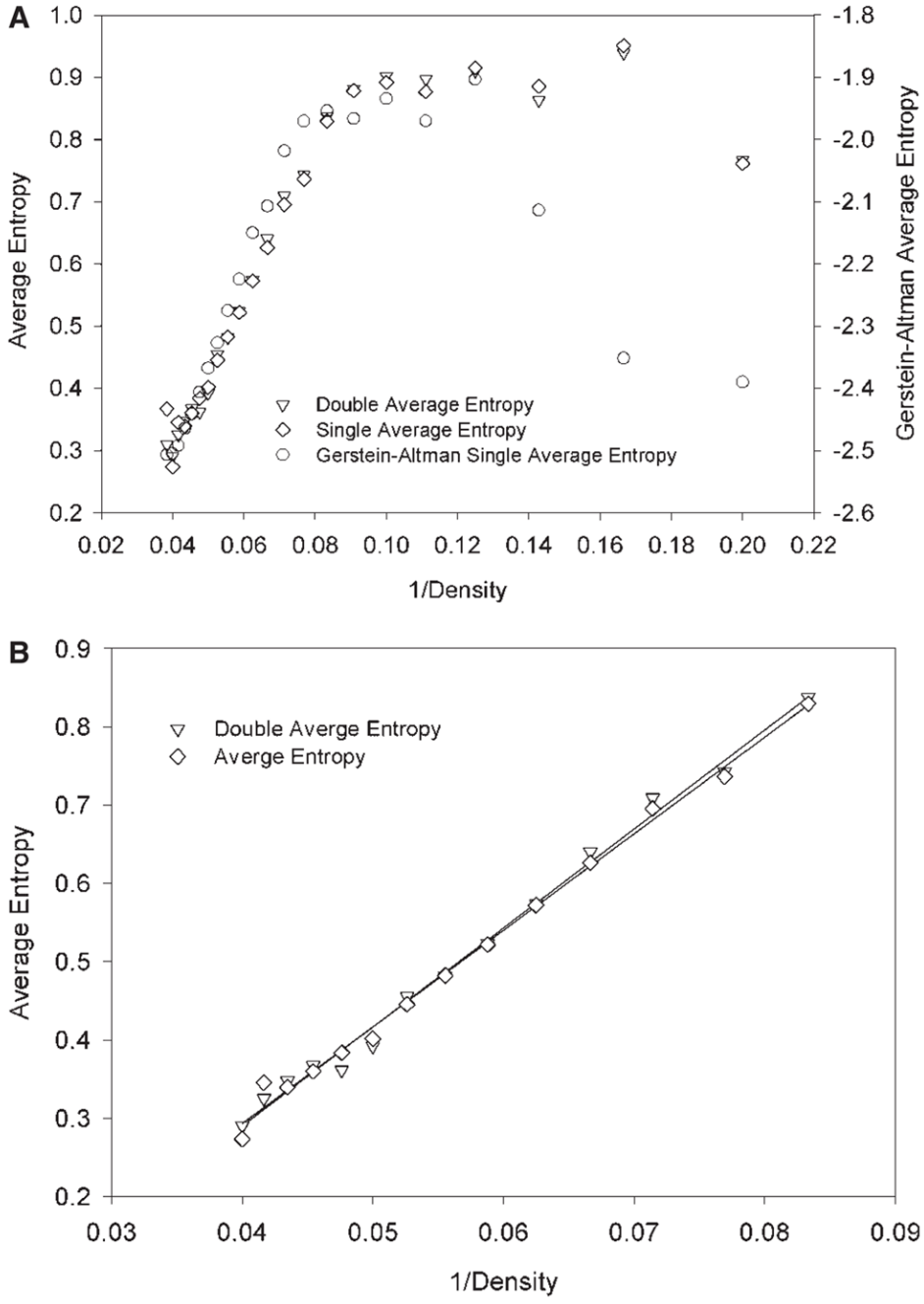

Fig. 3.

Correlation plots of overall average entropy for the set of 130 proteins with inverse packing density. (**A**) Inverse packing density (ordinate) is calculated from C^{α} packing density noted in Figure 2 and overall sequence entropy (ordinate) is calculated in three ways: single averaged (open diamonds) and corrected for randomness as noted on the right ordinate (open circles) and 'double' averaged (open triangles). Single averaged entropy is determined by averaging sequence entropy for each associated residue position within its interval of inverse of packing density (abscissa). The estimated standard deviation with and without corrections for randomness is 0.5. 'Double' averaged sequence entropy is calculated by first averaging each protein's sequence entropy for a particular density interval and subsequently averaging over

all proteins. The estimated standard deviation is 0.3. (**B**) Linear regression of the selected Region I with 31 169 averaged residue entropy values (ordinate) out of the total of 41 632 aligned query residues. These averaged sequence entropy values correspond to the region of inverse packing density (abscissa) between 0.040 and 0.083 (or 25 to 12 C^{α} atoms within a 9 Å radius). Overall single averaged entropy (open squares) is fitted with a straight-line $y = 12.350x - 0.20$ with correlation coefficient 0.997 and $P < 0.001$. The 'double' averaged entropy (open triangle) straight-line fit is $y = 12.658x - 0.22$ with correlation coefficient 0.997 and $P < 0.001$. Note that between 0.040 and 0.083 inverse packing density, the single averaged entropy corrected for randomness has a straight-line fit $y = 12.409x + 3.05$ with correlation coefficient 0.998 and $P < 0.001$.

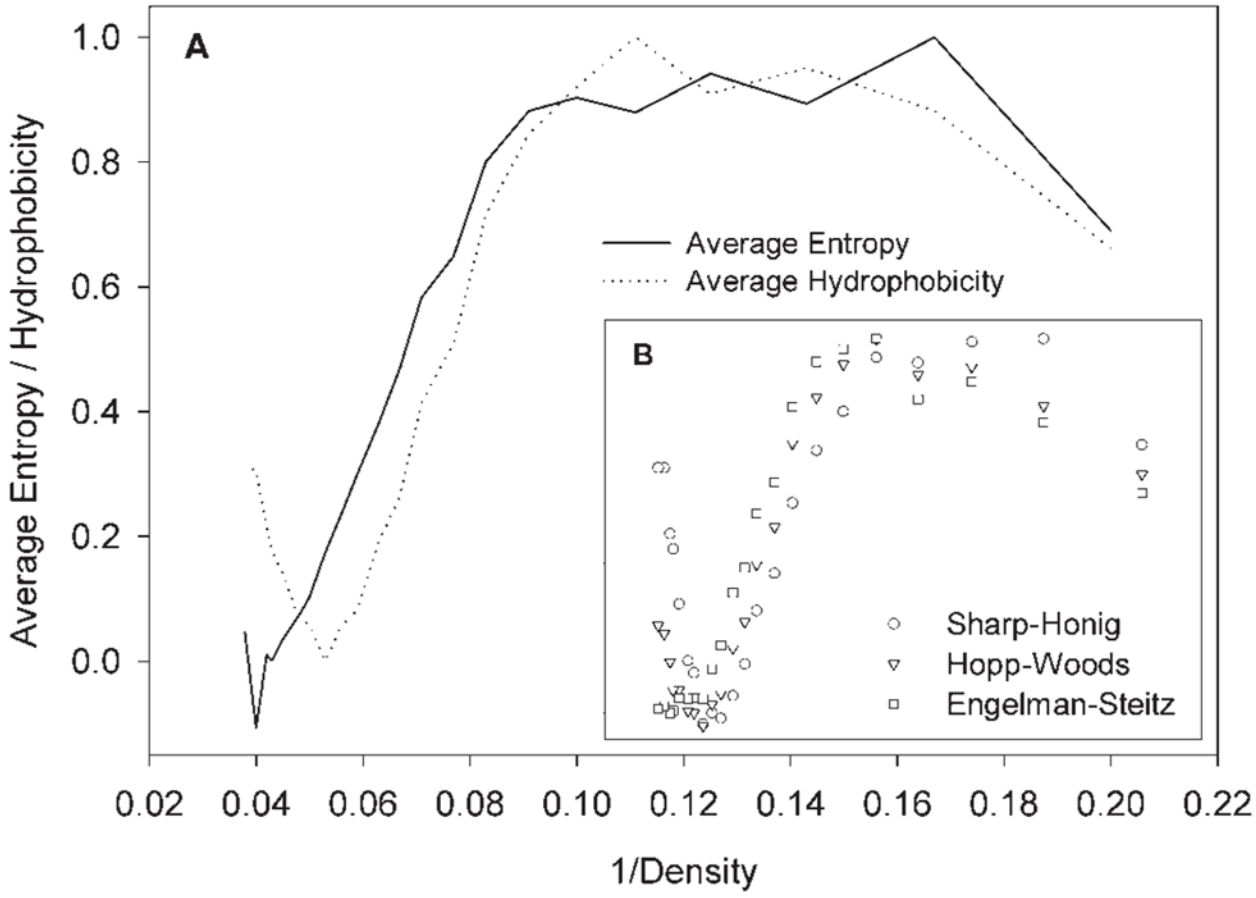

Fig. 4.

Comparison of average hydrophobicity per residue and overall single-averaged sequence entropy with respect to inverse C^{α} packing density. Residue hydrophobicity is calculated for each query protein, weighting the aligned residue type with the different scales: Hopp and Woods (1981), Engelman *et al*. (1986) and Sharp *et al*. (1991). The average hydrophobicity for each scale is calculated by averaging the residue hydrophobicities for all aligned residues within an interval of packing density. (**A**) Each of three sets of hydrophobicities corresponding to the different scales are normalized and then their average is renormalized (dotted line). The single-averaged sequence entropy from Figure 3A is normalized (solid line) and also plotted against inverse density. (**B**) Inset shows the corresponding three normalized sets of hydrophobicities plotted against inverse density, from Sharp *et al*. (diamonds), Hopp and Woods (inverted triangles) and Engelman *et al*. (squares).

Table I

List of 130 proteins

1a1i[a]	1agm	1aqh[a]	1bg3	1crc,[ab]	1eeh	3cla[b]	5acn
1a1s[ab]	1agx	1atp[a]	1bia[b]	1crm	1hgu,[ab]	3cna	5cha
1a32	1aha[b]	1av5	1bit[a]	1crz	1lz1[a]	3est	5cpa
1a3c[b]	1ahn	1av6	1blz	1csr	1omd[b]	3gbp	5cpv[b]
1a3s	1ai2[a]	1av7[a]	1bn6[b]	1d6m	1rbp[b]	3grs	5cts
1a48	1ak2[b]	1aw5[b]	1bo6	1daj	1rhd[b]	3pfk	5ldh[a]
1a59[b]	1ako	1aw9[b]	1boh	1dcs[b]	1ton	3pgk[a]	5rub[b]
1a5z[b]	1al8	1aye	1bsi[a]	1dhs	2act[a]	3pgm	6ldh[a]
1a6f[b]	1aln	1ayl	1bt3	1dht	2cts	3psg[a]	6xia
1a6q	1alc[b]	1ayx	1bul[a]	1din	2lbp	3rn3,[ab]	7api[a]
1aat	1amn	1azi,[ab]	1bxq	1dmr	2ldx[a]	3rp2	7cat[a]
1ab4[a]	1amp	1ba3[b]	1byt[b]	1e1k	2liv	4ape	8adh[a]
1acb	1an9	1bc2[b]	1cb0	1e3h	2prk	4dfr	8atc
1add	1ang[b]	1bf2	1cex	1e3q	2rn2	4mdh[a]	8dfr
1adi[a]	1ao5	1bfd	1cjx	1e5m[a]	2taa	4pep[a]	9pap[a]
1ae4	1aob[a]	1bg0	1ck6	1ebv	3blm	4tnc,[ab]	9wga
1af3[b]	1aq0						

[a] Indicates default maximum of 100 alignments.

[b] Proteins with *P* values in the range 0.001–0.15 (all others $P < 0.001$).